

Modeling and predicting page-view dynamics on Wikipedia

Marijn ten Thij
University of Twente
the Netherlands

Yana Volkovich
Barcelona Media Foundation
Spain

David Laniado
Barcelona Media Foundation
Spain

Andreas Kaltenbrunner
Barcelona Media Foundation
Spain

ABSTRACT

The simplicity of producing and consuming online content makes it difficult to estimate how much attention will be devoted from Internet users to any given content. This work presents a general overview of temporal patterns in the access to content on a huge collaborative platform. We propose a model for predicting the popularity of promoted content, inspired by the analysis of the page-view dynamics on Wikipedia. Compared to previous studies, the observed popularity patterns are more complex; however, our model uses just few parameters to fully describe them. The model is validated through empirical measurements.

Categories and Subject Descriptors

H.5.3 [Information Interfaces]: Group and Organization Interfaces—*Computer-supported cooperative work, Web-based interaction*

General Terms

Human Factors, Measurement, Theory

Keywords

Wikipedia, promoted content, temporal patterns, popularity prediction

1. INTRODUCTION

The social media boom gave a birth to a wide range of studies about online traces generated by Internet users. One of the important research targets addressed by these studies is the analysis and prediction of the dynamics of content popularity. Historically, most of the studies were focusing

on the analysis of content generated on blogging [3, 7], later microblogging [9], video-sharing [19] and news-sharing platforms [5]. However, in many cases the studies reflect only the behavior of registered users or focus on a website of interest only for a specific community, e.g. Slashdot [5]. Here we analyze instead a website of general interest and address the problem of understanding online usage and popularity patterns through a large-scale analysis of the visitors and the users of Wikipedia, the sixth most visited website¹.

Wikipedia is a free, collaboratively edited and multilingual Internet encyclopedia. It has an estimated number of 365 million monthly readers worldwide². Although many studies looked on editing and commenting activity on Wikipedia [6, 13, 17, 22], there are not many quantitative works focusing on the Wikipedia usage by the Internet users. To the best of our knowledge, there are just few studies which explore Wikipedia views as an information source in order to detect and predict events in real world. Osborne et al. [14] used a stream of Wikipedia page views to improve the quality of discovered events in Twitter, and Mestyan et al. [11] predicted the popularity of a movie by measuring the activity level of editors and viewers of the corresponding Wikipedia entry. Finally, in [15, 16] authors analyzed how the Wikipedia traffic data is influenced by external and internal events.

One of the main goals of this work is to examine content popularity on Wikipedia. Similar to many online platforms, on Wikipedia some of the articles get promoted to the *Main page*³. In Figure 1 we present an example of the *Main page* of the English Wikipedia. Every Wikipedia user can nominate any article to the pool of possible future promoted articles (*featured articles*) on a specific page⁴.

We refer to the article placed under the headline “From today’s featured article”⁵ as *promoted*, in order to avoid confusion with featured articles in general. Every day, at 00h(UTC), a new promoted article is placed on the *Main page* together with links to the three articles promoted during the previous three days (see “Recently featured” at the bottom left of Figure 1). The “today’s promoted” articles are also sent by e-mail to subscribers at 04h(UTC).

¹<http://www.alexa.com/siteinfo/wikipedia.org>

²S. West, “Wikipedia’s Evolving Impact” (TED2010) <http://goo.gl/erGp2>

³http://en.wikipedia.org/wiki/Main_Page

⁴<http://en.wikipedia.org/wiki/WP:FAC>

⁵http://en.wikipedia.org/wiki/Wikipedia:Today's_featured_article



Figure 1: English Wikipedia *Main page* on December, 20, 2012.

The simplicity of producing and consuming online content makes it difficult to predict how much attention will be devoted from Internet users to any promoted item. On one hand, a number of studies [19, 20] tries to address this question by analyzing media-sharing platforms such as Digg or Youtube. However, such platforms rank and categorize content based on previous popularity and user votes, and this leads to rich-get-richer bias in the number of views and in the duration of the promotion time. On the other hand, the concept of content promotion on Wikipedia is distinctively different. Promoted articles on Wikipedia are generated and managed through online collaboration and shown to the online audience for a fixed amount of time. This predefined exposure duration makes the Wikipedia promotion to have in a way more in common with online advertisement than with content popularity on other media platforms.

The rest of the paper is organized as follows. In the next section we briefly discuss our main contributions. Then, in Section 3 we describe data sets we use in this study. In Section 4 we focus on the general statistics of Wikipedia traffic and compare the article-view data with the editing and commenting activities. In the same section we also look on the average promoted article popularity. Next, in Section 5 we introduce the model to describe the number of views during the exposure of a promoted article. In Section 6 we use this model to predict the number of views during the exposure time of an article. Finally, we discuss the related studies in Section 7 and present our conclusions in Section 8.

2. OUR CONTRIBUTION

In this work we aim to explore temporal and popularity patterns on the English Wikipedia. In particular, our primary interest lies in the number of views a page receives per hour. We first focus on the overall interest of the Internet users on Wikipedia. We analyze total view statistics for the English Wikipedia, and argue that these values show some tendency to daily and weekly cycles. Second, we compare the number of the Wikipedia views with the number of comments and edits. The temporal characteristics for the latter two measures have been analyzed in [6]. We observe that the overall dynamics can be described as “There are more and more readers of Wikipedia, but they have less and less new to add” [4].

Moreover, given a predefined set of pages, which we will describe in Section 3, we analyze the daily correlations between views, comments, edits and distinct editors. We find

low but positive correlations when only articles with non-zero activities are considered.

Finally, we identify a specific popularity pattern for Wikipedia content, in particular, for the number of views of “today’s featured articles”. We introduce a model to describe this pattern and use this model to predict the popularity of promoted Wikipedia content. The model should be applicable to analyze and compare popularity patterns of promoted content on collaborative platforms in general.

3. DATASET

We retrieve the page-view values from a database provided by *Wikimedia*⁶. In this database we find a file, for every hour, which lists the total number of views to a page during that hour, provided it received at least one view. We extract the page view data between December 9, 2007 at 18h(UTC) and March 31, 2010 23h(UTC), for a total of 844 days. Note that this database is not entirely complete: for some hours there is missing data (see Table 1 for the largest time gaps) or there is more than one entry for a single article. For the latter we just sum these entries. Finally, we assume that views to articles which come through redirects are also registered in the data of the target page.

Table 1: Missing data in page views (UTC).

Date	Missing hours
March 3 - March 4 2008	18h - 16h
August 21 - August 22 2008	12h - 12h
September 21 - October 1 2009	17h - 0h
October 15 - October 16 2009	0h - 2h
January 23 - January 25 2010	3h - 1h

We also used the Wikipedia dump from March 12, 2010 (see [8] and [6] for more details) to extract the temporal data of edits and comments for the Wikipedia articles. Combining these datasets we select 871 395 articles that have been commented on at least once in their history. These articles accumulated $32 \cdot 10^9$ views in total which is on average $38 \cdot 10^6$ views per day.

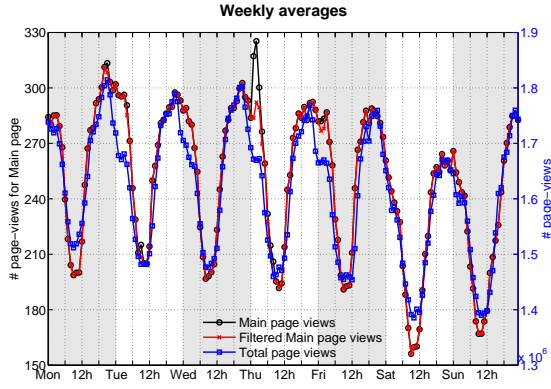
4. PAGE VIEW STATISTICS

We start with the description of the hourly and daily number of views of the English Wikipedia. Next, we compare the historical trends in the page views with the trends in the number of edits and comments. Finally, we speak about the popularity of a promoted article during the promotion period.

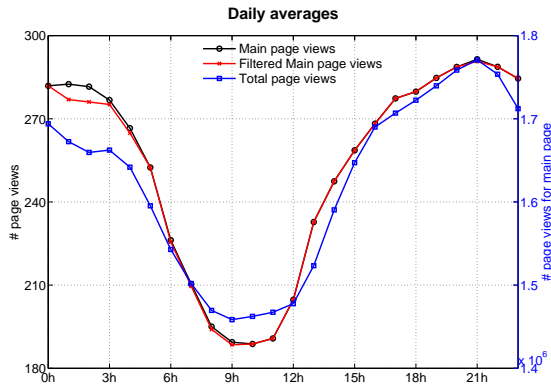
4.1 Circadian and weekly patterns

In Figure 2 we observe the number of visits per hour to the English Wikipedia in general and also to its *Main page* averaged by weeks and by days. We observe that the number of views of English Wikipedia varies between $1.4 \cdot 10^6$ and $1.8 \cdot 10^6$ with an average of $1.6 \cdot 10^6$ views per hour. For the *Main page* popularity we find that there are in average 252 visitors per hour. Interestingly, in Figure 2(a) we observe a significant jump in the number of *Main page* views from 01h to 03h(UTC) on Thursdays, that can be explained by the presence of some extreme outliers in the *Main page* views in our dataset. If we remove these outliers (red curve in Figure 2(a)) we observe a weekly pattern similar to the

⁶<http://dumps.wikimedia.org/other/pagecounts-raw/>



(a) weekly patterns



(b) circadian patterns

Figure 2: Temporal patterns of the number Wikipedia views per hour in total and for its *Main* page.

overall weekly activity which is slightly decreasing during weekends.

Figure 2(b) depicts the circadian patterns of the Wikipedia page views. We observe the lowest activity between 09h and 11h(UTC) corresponding to the night hours in the US (taking US Central Time (UTC-6)). Similar circadian patterns but for editing activity on English Wikipedia were observed in [22]. In the following section we discuss the relations between editing, commenting and viewing activities in more detail.

4.2 Views, edits and comments

In [6] Kaltenbrunner and Laniado analyzed changes in the global number of edits and comments in the English Wikipedia. In particular, they confirmed the decreasing trend in editing and commenting activities [17] and obtained the ratio of 6 comments per 100 edits. In Figure 3 we use this ratio to display the global number of edits and comments per day on one scale. In the same figure we also plot the corresponding global number of daily views. We note that in the latter case the gaps in the curve correspond to the data gaps (see Table 1). We observe that the global number of views of Wikipedia grows over time. Similar results were obtained in [1], where the authors analyzed the view-trends

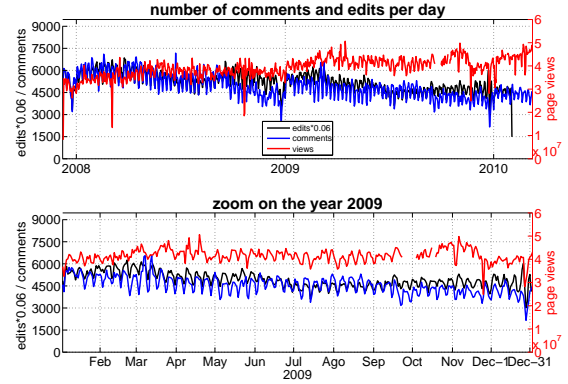


Figure 3: Evolution of the page views, comments and edits over time.

for some selected page-categories. Therefore, the trend of using Wikipedia is growing in the same time as the trend of making Wikipedia is lessening.

Our next step is to look on the correlations between the global number of views, edits and comments per day. To this end we calculate Spearman's rank correlation coefficients (see Table 2). Here we use Spearman's coefficient because the values under analysis demonstrate heavy-tail behavior and also possess values on different scales. We refer to [10] for more arguments on why it is essential to use rank correlations for heavy-tailed distributed data. In Table 2 we see that the global number of views per day and the corresponding number of edits and comments are negatively correlated. This observation can be explained by the differences in trends for these parameters, in particular, the global number of article-views is increasing, while other parameters are decreasing. If we exclude these trends from the data by removing the best linear fit, we find that the number of views indeed is correlated with the number of edits and comments (values in brackets in Table 2). Interestingly, removing the linear trends decreases the value of correlation coefficient between the number of edits and comments. Though, these characteristics remain very correlated.

Table 2: Spearman's rank correlation coefficients for the global number and the detrended global number (in brackets) of views, edits, and comments.

	# comments	# edits
# views	-0.24 (0.40)	-0.22 (0.43)
# comments	-	0.75 (0.53)

Finally, we look on the relation between the number of views and the number of edits and comments per article separately. To this end, for each day we calculate the corresponding correlation coefficients for the number of views, edits, comments, and distinct editors based on the values observed for the pages from a specific article set. In the *complete* set we include all articles that have received at least 1 comment in their history. Then, for each day and for each pair of characteristics we also construct *OR* and *AND* sets as follows: we take only pages which have at least one or both non-zero values for the selected characteristics for the selected day. We find that these daily correlation coeffi-

Table 3: Spearman’s rank correlation coefficients for the number of views, edits, and comments an article receives per day (*complete/OR/AND* sets).

	# edits	# editors	# comments	set
# views	0.20	0.20	0.04	<i>complete</i>
	0.19	0.19	0.04	<i>OR</i>
	0.29	0.36	0.16	<i>AND</i>
# edits	-	0.74	-0.24	<i>complete</i>
	-	0.74	0.23	<i>OR</i>
	-	-	0.07	<i>AND</i>
# editors	-	-	-0.28	<i>complete</i>
	-	-	0.21	<i>OR</i>
	-	-	0.21	<i>AND</i>

cients are quite stable over time (plots are not shown) with the average values reported in Table 3. We observe that in spite of the fact that every edit and comment implies a view, the number of views per article during a selected day is only weakly correlated with the number of edits or editors. The correlation between the number of views and comments is even lower. However, the correlation is strongest for the *AND* set in all three cases. This indicates a stronger connection between these quantities for articles with larger edit and comment activity. Interestingly, we observe a negative correlation between the number of comments and edits per day for the *OR* set. This may indicate that comments and edits are made at different times in articles with low activity.

4.3 Promoted articles

In the previous sections we have analyzed the temporal characteristic of activities on Wikipedia articles in general. In the rest of this work we will focus on the page-view data for the promoted articles only. Recall that in Wikipedia an article gets promoted for a predefined period of 1+3 days, which we call the exposure duration in analogy to [20]. We restrict our analysis and predictions only to these article exposure durations.

We select all the articles promoted in the time-span from January 1st, 2008 through March 31st, 2010, which is a total of 822 articles. Among them only 686 have complete page-view data, i.e. we know the number of views for every hour in their exposure duration. We also omit two more articles

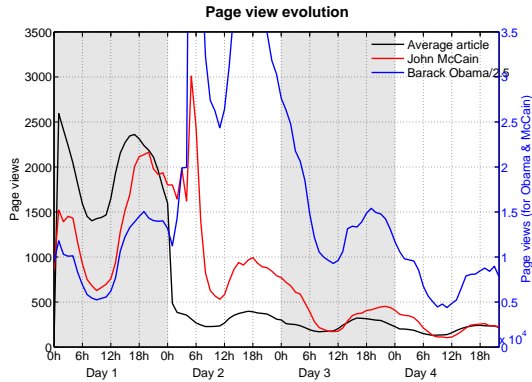


Figure 4: Progression of the number of views for the promoted articles *Barack Obama* (divided by 2.5 for scaling), *John McCain*, and average article during their exposure time.

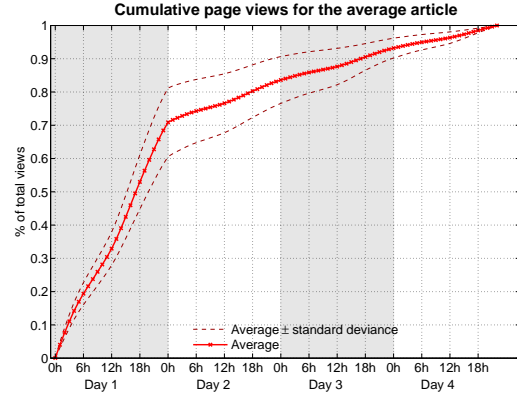


Figure 5: Average normalized popularity of the promoted articles during the exposure period.

(*Barack Obama* and *John McCain*) which have been both promoted on November, 4th, 2008⁷. These articles, with the largest number of views during the second day of exposure, show completely different dynamics (see Figure 4) compared to the average article (black curve). These pages, therefore, would influence some of the results reported below. Thus, in this study we use only 684 promoted articles.

By popularity of a promoted article we mean the number of views this article receives during the exposure duration. In Figure 5 we show the average normalized popularity of a promoted page. Thus, for every promoted article we first find the total number of views this article receives at the end of the fourth day and then we use this number to obtain time series of popularity that monotonically increase from 0 at the moment of the article promotion to 1 at the end of the exposure duration. Using such approach allows us to eliminate the differences in specific interestingness among the articles. We observe a clear difference in dynamics of popularity during the first and the rest of the exposure days. In both cases we see an approximately linear increase of popularity which is very different with strictly-concave results for Digg and Youtube reported in [19].

In Figure 6 we plot the average number of views a promoted article attracts during the t -th hour v_t , $t = 1, \dots, 95$. We clearly see that the exposure period of a promoted article in Wikipedia can be divided into four stages. At the first stage or at the first hour after a page gets promoted, we witness a huge increase in the article’s popularity. The value v_1 even is the largest for the average promoted article. The second stage contains the remaining hours of the first day of the promotion. The third stage is characterized by the negative jump occurring after the original article gets replaced by the new one. Finally, the last stage contains the view dynamics during the 3 days of being promoted in “Recently featured”. Using this stage-representation, we construct $g(t)$ as a piecewise-linear approximation of $\log(v_t)$ and plot it in Figure 6. We refer to Appendix for more details.

Comparing approximation $g(t)$ and the promoted article popularity v_t we notice that the main differences are caused by the circadian patterns of Wikipedia views. In order to re-

⁷This is the only occasion where 2 articles are promoted at once for the reason of the US presidential elections

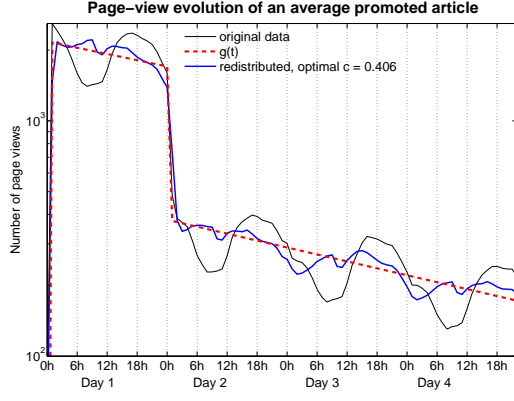


Figure 6: The average number of views of a promoted article arranged by different time scales.

move this variation we use a new time scale in which every hour is measured in the number of views rather than in minutes. This approach was introduced in [19] for Digg stories popularity. We modify the original idea by removing a constant fraction c of the traffic data to emphasize the circadian patterns even more. In the rest of paper we call the new time scale as *redistributed time scale* and set one hour in this scale to be equal to $1/24$ of the product of $(1 - c)$ and the average number of *Main Page* views per day. We again refer to Appendix for more details. In Figure 6 we plot the average number of views of the promoted articles in the redistributed time scale ($c = 0.406$) and observe the linear decreasing trend of the promoted article popularity with time.

5. MODEL

Based on the view-behavior of the average over all promoted articles we propose a model which completely describes the traffic dynamics of a selected promoted article during its exposure duration. This model is defined by two parameters: a constant interest-decay factor for all days of the promotion and negative jump of the popularity after the first day of the exposure. The number of views the selected article receives during the first hour of the promotion is used as the only input value of the model.

5.1 Model definition

In this section we define the model which explains the distribution of the number of views per time unit of a promoted article during the exposure duration. We start by describing the specific shape of this distribution in redistributed time (removing the circadian cycle). To this end we set $\hat{w}_{1^*} := 1$ and define the normalized number of views for the promoted article at any redistributed time t^* , $t^* = 2, \dots, 95$, as

$$\hat{w}_{t^*} = \beta_{t^*} \cdot \hat{w}_{t^*-1}.$$

Previously, we have described the four stages of the exposure life of a promoted article on Wikipedia. Using these definitions we set a temporal factor β_{t^*} as $\beta_{t^*} = \gamma$ for $t^* = 25$ and $\beta_{t^*} = \beta$ for other t^* 's, i.e. for $2 \leq t^* \leq 24$ and $26 \leq t^* \leq 95$. The constant factor β models the decay of the number of page-views in a typical hour of the exposure

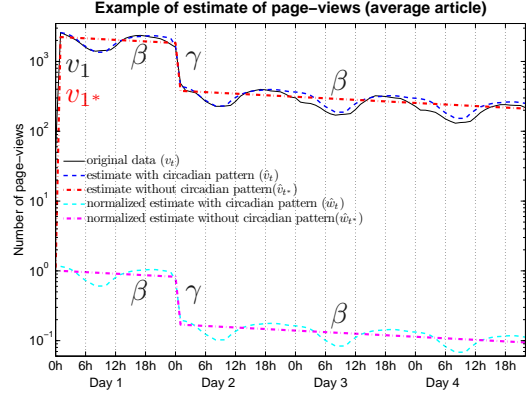


Figure 7: Explanation of the promoted article popularity model.

duration, i.e. while the page is promoted on the *Main page*. The factor γ states for the negative jump in the number of views after the promoted article gets moved to “Recently featured” position. Thus, we model the shape of the article popularity by stage: the first stage of the promoted article is characterized by \hat{w}_{1^*} , the second by interest-decay factor β , the third by γ , and the fourth again by the same factor β . To summarize, we model the normalized number of views of the promoted article during the t^* -th redistributed hour as follows:

$$\hat{w}_{t^*} = \begin{cases} \beta^{t^*-1}, & \text{for } 2 \leq t^* \leq 24; \\ \gamma \beta^{t^*-2}, & \text{for } 25 \leq t^* \leq 95. \end{cases}$$

Finally, we use the reverse time-redistribution to find \hat{w}_t , i.e. the corresponding number for \hat{w}_{t^*} but in the original time scale. We define the number of views of the promoted article during t -th hour as

$$\hat{v}_t = v_{1^*} \cdot \hat{w}_t,$$

where v_1 and $v_{1^*} = v_1/\hat{w}_1$ are the numbers of views of the promoted article after the first hour of exposure in the original and redistributed time scales. In Figure 7 we draw an explanation for the model.

The introduced model uses only the number of views during the first hour of the exposure period v_1 as an input parameter. In Figure 8 we plot the histogram for v_1 's in our dataset together with the log-normal fit ($\mu = 7.63$ and $\sigma = 0.71$). In the next section we focus on the estimation of parameters β and γ which are defining our model.

5.2 Estimation

We estimate the model's parameters β and γ by using page-view data of the promoted articles on the English Wikipedia. We apply the estimation algorithms for two sets of promoted articles. Thus, the first set S_1 contains all 684 articles and the second S_2 the first 100 promoted articles ordered by the date of promotion. We use S_1 in order to describe the general view dynamics for promoted content on Wikipedia. We use S_2 in Section 6 to predict the popularity of the remaining 584 promoted articles given the estimated β and γ and the corresponding initial values v_{1^*} 's.

We denote as $\hat{v}_{t^*}(\beta, \gamma)$ the predicted number of views for given values of β and γ at redistributed time t^* , and as

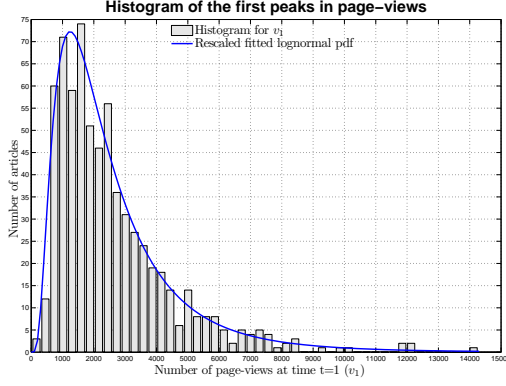


Figure 8: Histogram of occurring values for v_1 .

v_{t^*} the actual number of page views at time t^* for some promoted article $s \in S$, where S is either S_1 or S_2 . Then, we calculate parameters β and γ that minimize the error:

$$\{\beta, \gamma\} = \arg \min \sum_{s \in S} \left[\sum_{t^*=1}^{95} [\log(\hat{v}_{t^*}(\beta, \gamma)) - \log(v_{t^*})]^2 \right].$$

This yields to $\beta = 0.9915$ and $\gamma = 0.2805$ for S_1 and to $\beta = 0.9960$ and $\gamma = 0.4174$ for S_2 .

So far we have assumed that γ can be modeled as a constant factor for all articles. However, since γ encodes the negative jump in the decay of user interest after the first day of the exposure, we suggest that it should be correlated with the overall popularity of the promoted article. To this end, we compared γ with the total number of views a promoted article receives during one day before the promotion date but found no correlation between them (data not shown). Then, we propose to define γ as function of the initial popularity v_1 . We first find that $\log(v_1)$ and $\log(\gamma)$ are negatively correlated (Pearson's correlation coefficient is -0.31 for S_1 and -0.32 for S_2), which is also indicated in Figure 9. Then, we derive a log-linear function for γ :

$$\gamma = C \cdot v_1^m,$$

based on the observations $\{\log(v_1), \log(\gamma)\}$ for the articles from set S , where S is again either S_1 or S_2 . We rewrite the last equation in the following form:

$$\log(\gamma) = h(v_1) = m \cdot \log(v_1) + \log(C). \quad (1)$$

Using the set S_1 we obtain $m = -0.29$ and $C = 2.09$ for all articles. We note that for estimation of parameters m and C we omit the outliers⁸ indicated as red squares in Figure 9. We also perform the fitting for (1) on S_2 and obtain $m = -0.28$ and $C = 1.90$. Note that, although the initial estimates for γ were very different for S_1 and S_2 , the parameters of $h(v_1)$ are not. This can be also observed in the nearly overlapping linear fits in Figure 9. This may be explained by different mean initial views v_1 's in S_1 and S_2 .

Comparing the estimated values for $\log(\gamma)$ with $\log(h(v_1))$, we find that $\log(\gamma) \sim \mathcal{N}(h(v_1), \sigma^2)$. Therefore, we can derive

⁸These are the articles *Borobudur*, *Princess Beatrice of the United Kingdom*, *Local Government Commission for England (1992)*, *West Indian cricket team in England in 1988* and *Attachment theory*.

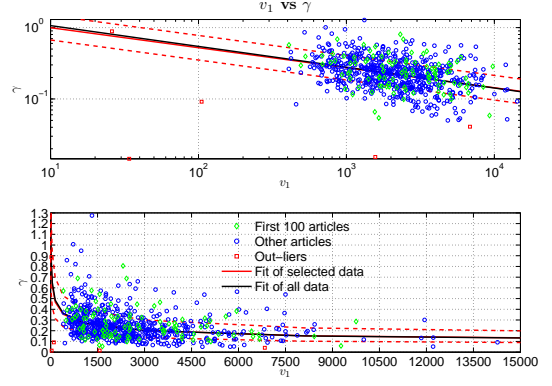


Figure 9: Relations between γ and v_1 . The dashed red lines indicate the interval $[h(v_1) - \sigma, h(v_1) + \sigma]$.

an interval in which the decay factor would lie with a given probability. We use $[h(v_1) - \sigma, h(v_1) + \sigma]$ as this interval, indicated by the dashed lines in Figure 9.

Back to the model, we can now calculate \hat{w}_{t^*} at time $t^*, t^* = 2, \dots, 95$ as follows:

$$\hat{w}_{t^*} = \begin{cases} \beta^{t^*-1}, & \text{for } 2 \leq t^* \leq 24; \\ C \cdot v_1^m \cdot \beta^{t^*-2}, & \text{for } 25 \leq t^* \leq 95; \end{cases} \quad (2)$$

and then use the reverse time-redistribution to find \hat{w}_t . Using \hat{w}_t and

$$\hat{v}_t = \frac{v_1}{\hat{w}_1} \cdot \hat{w}_t \quad (3)$$

we can obtain the estimated hourly progression \hat{v}_t of the page-views for $t = 2, \dots, 95$.

6. POPULARITY PREDICTION

As we have already discussed before, we use the first 100 promoted articles (ordered by the date of their exposure) to learn parameters β and γ in order to apply our model for prediction of the popularity of the Wikipedia content for the remaining 584 promoted articles from our dataset. Thus, for each of these articles we take the article's popularity after the first hour v_1 and use Equations (2) and (3) for parameters $\beta = 0.996$ and γ ($m = -0.28$ and $C = 1.9$) from the previous section.

As we will discuss below for most of the promoted articles we are able to obtain a good prediction for the page-view dynamics during the first day of exposure. However, for the remaining days the number of actual page-views v_t does not always lie within the predicted interval $[h(v_1) - \sigma, h(v_1) + \sigma]$ for $t = 25, \dots, 95$, as we see in Figure 10. Thus, although in the 25-th hour we correctly predict the page popularity for 74% of the articles, in general we observe a decreasing trend in the percentage of the correct predictions. This is caused by underestimating the decline of interest (or an overestimation of γ) by our model and can be improved by introducing the input parameter v_{25} , i.e. the value of the promoted page popularity right after it is moved to the "Recently featured" section, into our model.

Adjusting the prediction during the first hour of the second day of the promotion with v_{25} leads us to the following

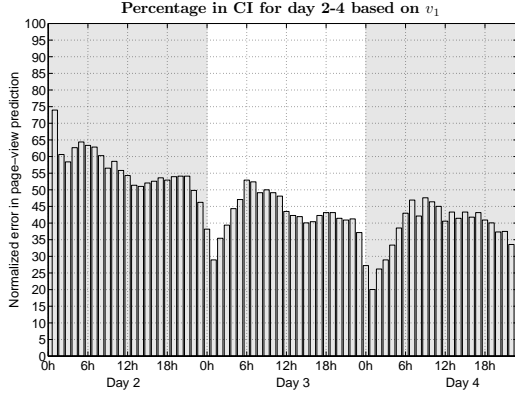


Figure 10: Percentage of values v_t in predicted interval for $t = 25, \dots, 95$.

description of the model:

$$\hat{w}_{t^*} = \begin{cases} \beta^{t^*-1}, & \text{for } 1 \leq t^* \leq 24; \\ \beta^{t^*-25}, & \text{for } 25 \leq t^* \leq 95. \end{cases}$$

Here we again use the reverse time-redistribution to find \hat{w}_t to obtain the predicted hourly page-views progression \hat{v}_t , for $t = 2, \dots, 95$, by calculating

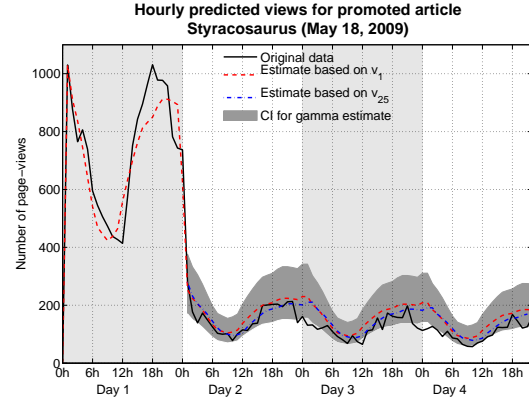
$$\hat{v}_t = \begin{cases} \frac{v_1}{\hat{w}_1} \cdot \hat{w}_t, & \text{for } 1 \leq t \leq 24; \\ \frac{v_{25}}{\hat{w}_{25}} \cdot \hat{w}_t, & \text{for } 25 \leq t \leq 95. \end{cases}$$

In Figure 11 we present two examples for the prediction of the popularity for both of the above-defined prediction methods. We show the initial prediction in red, the interval $[h(v_1) - \sigma, h(v_1) + \sigma]$ as dark grey area and \hat{v}_t based on v_1 and v_{25} in blue. While the prediction of the article *Styracosaurus* performs well already using only v_1 , similar prediction overestimates the views of the article *Alice in Chains*.

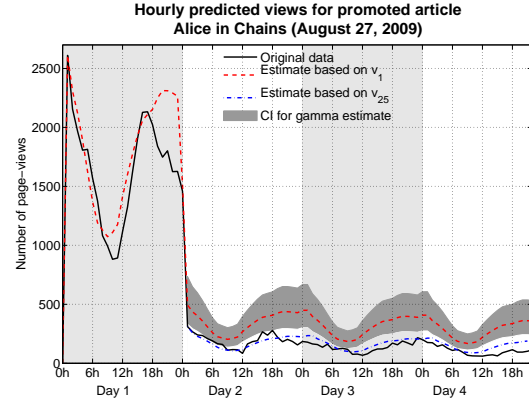
We analyze the normalized hourly errors $\left(\frac{\hat{v}_t - v_t}{v_t}\right)$ for all articles under study for both prediction methods: errors for just v_1 are plotted in red, while errors using both v_1 and v_{25} in blue. From Figure 12 we observe that our prediction performs well for the first day of exposure. We recall that for this time interval we only use v_1 for the prediction. For the second, the third and the fourth days we observe an increase of the spread of hourly errors. However, this increase is much smaller for the second prediction technique.

In Figure 13(a) we present the distribution of the maximum normalized hourly error for predictions: both are right-skewed. For the method which only uses v_1 as an input we observe more large overestimates. The distribution of the minimum (maximum negative) normalized hourly error is displayed in Figure 13(b). We observe that it is approximately normally distributed for both prediction methods, but showing larger underestimates for the method using v_1 and v_{25} .

Finally, we present the absolute hourly errors $(\hat{v}_t - v_t)$ in Figure 14. Interestingly, we observe that the absolute error towards the end of day 1 is large. This is caused by the fact that we model the decay only to be during one specific hour whereas for most articles it actually starts a few hours before the end of the first day of the exposure duration. We also see that the hourly error during the second, the



(a) *Styracosaurus*



(b) *Alice in Chains*

Figure 11: Examples of the prediction of the page-views for a promoted article.

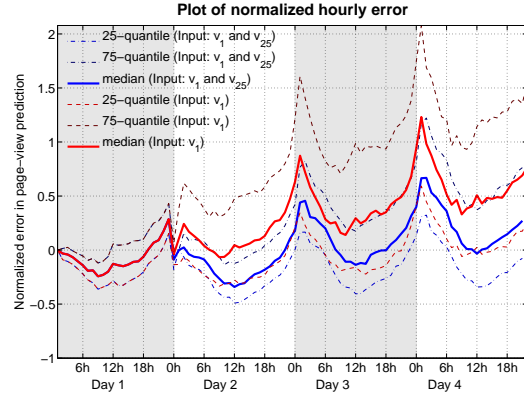
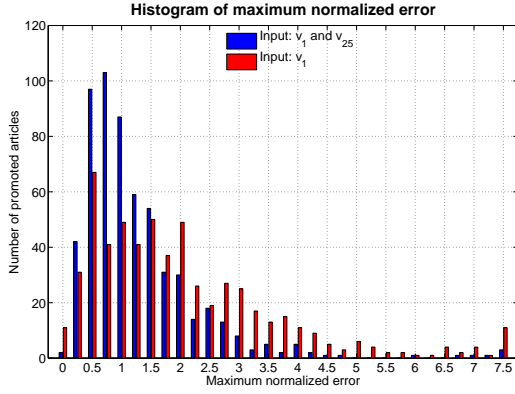
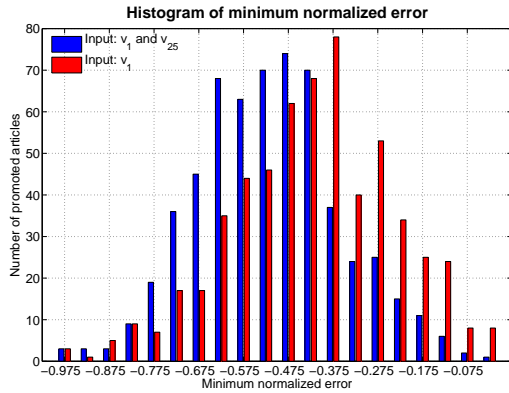


Figure 12: Hourly normalized errors.

third and the fourth days are slightly increasing and follow a circadian pattern. This is similar to the observation in Figure 12. Again, the prediction method which uses both v_1 and v_{25} as input outperforms the model that only uses v_1 .



(a) maximum normalized error



(b) minimum (maximum negative) normalized error

Figure 13: Distribution of the normalized maximum and minimum errors of actual page-views.

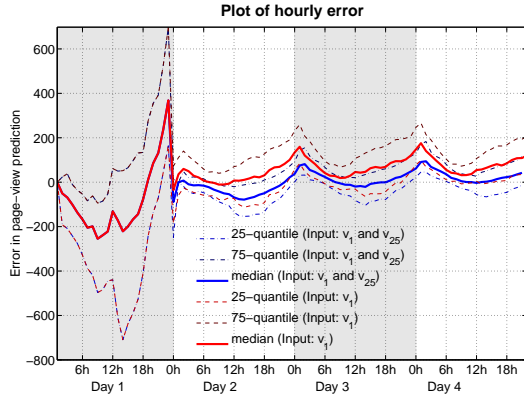


Figure 14: Hourly errors of actual page-views.

7. RELATED WORK

There is a vast literature analyzing different aspects of Wikipedia, e.g. see [12] for an overview. In this section we discuss only studies that are closely related to our work. Thus, in [16] authors provided a high-level overview of Wiki-

pedia traffic for 2009 with a particular focus on content-type. In [15] Ratkiewicz et al. analyzed the Wikipedia page-view data for the thirteen month span in 2008–2009. In particular, they reported the heavy-tailed distribution for the number of views per page. The authors also argued that the top bursty articles in their dataset, where the “burstiness” was defined as the ratio of the article’s traffic of its present to previous day, can be divided into two sets. In the first set the articles traffic was influenced by the external events and correlated with Google Trends results. In the second set the pages accreted their traffic due to internal Wikipedia dynamics and, in particular, this traffic was correlated with the hits neighbors of these pages received. In [11] authors explored relations between the external factors and the Wikipedia traffic by using the article-view data for the movies to predict the opening box office takings of blockbuster movies.

In [6] Kaltenbrunner and Laniado analyzed the peaks in editing and commenting of *Barack Obama* article on Wikipedia with the corresponding political and social events. They also found that some of the peaks in these activities were due to some internal Wikipedia dynamics. The circadian patterns for editing behavior for the different language Wikipedias were recently investigated in [22].

A number of studies focused on promoted content and news popularity. In [2] authors reported power laws in the distribution of views of the short-lived events, such as news on a large Hungarian news portal. They also found that for most of the news items, the number of views decays significantly 36 hours after posting.

The log-normal distribution was found to describe the number of comments per minute a news story on Slashdot receives [5]. The authors combined this fact with the circadian activity cycle of the website to be able to predict the number of comments a news post on Slashdot receives per time unit. A similar idea based on rescaling average access patterns was also used in [18] to predict the long time popularity of content on Youtube and Digg. The later study introduced the notion of Digg-time as we use in here in this study.

In [21] Wu and Huberman confirmed the log-normal nature of the news story popularity on the social news portal Digg. They also proposed a stochastic dynamic model which defines the page-view increase as a random variable multiplied by some converging to zero factor and the previous popularity. The expected value of the random variables modeled the fraction of people which would spread the news story to their social neighbors. The stochastic model was later expanded for Reddit and Epinions [20] and Youtube [19], extending their work of [18]. Similar to our model, they used a single value as the baseline for their prediction and they used a parameter to model the decay of interest. However, in contrast to our model, they added an extra random variable, to account for randomness in the data they extracted.

8. CONCLUSIONS

We have presented a simple yet powerful model for the view dynamics of promoted content on Wikipedia. The model shows that the number of views an article receives decays exponentially in time with a constant decay rate, if the dependency of the data on Wikipedia’s circadian activity cycle is removed. The only exception from this decay rule

is the presence of a larger decay when an article is moved from the “today’s featured” to the list of “Recently featured” after 24h of being promoted. The model allows to predict the popularity of an article using only the number of views it received during the first hour of exposure. The quality of the prediction can be improved if the model is updated right after an article is moved to “Recently featured” section.

Our model should allow to describe and compare view dynamics on other websites or parts of websites with similar update strategies, e.g. online newspapers which are updated on a daily basis, or a list of today’s recommended items (mobile apps, products, etc). The decay factor might be a useful parameter to account for the half-live of a piece of content on a given site. The findings might also be useful to predict the success rate of new online advertisements or sponsored content in general.

In this study we focused only on promoted articles, which provide a nearly ideal experimental setup to study content popularity. There is no competition between articles for the users’ attention as there is only one promoted article per day. Also the time of promotion is periodical and fixed which gives all articles the exact same attention time spans. However, this is not the only section on the *Main page* that can be investigated. Future work should extend the model to other sections with more complicated settings like *In the news*, *On this day* and *Did you know*.

Furthermore, the occurrence of page-view peaks should be investigated. These findings could be compared to the results of peaks in editing and commenting behavior [6]. This would give more insight into patterns of sudden attention spikes of users in Wikipedia.

Finally, we have observed a growing trend in page views on the English Wikipedia, whereas the amount of edits and comments is decreasing. Thus it seems that, indeed, there are more and more readers of Wikipedia, but it becomes increasingly more difficult to add something new.

9. ACKNOWLEDGEMENTS

The authors gratefully acknowledge Nelly Litvak for her helpful comments and suggestions. Yana Volkovich acknowledges support from the Torres Quevedo Program from the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund.

10. REFERENCES

- [1] A. Capiluppi, A. C. D. Pimentel, and C. Boldyreff. Patterns of creation and usage of Wikipedia content. In *Proceedings of WSE 2012*, 2012.
- [2] Z. Dezső, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A. Barabási. Dynamics of information access on the web. *Physical Review E*, 73(6):066132, 2006.
- [3] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *Proceedings of ICWSM-06*, 2006.
- [4] R. Jensen. Military history on the electronic frontier: Wikipedia fights the war of 1812. *Journal of Military History*, 76(4):1165–82, 10 2012.
- [5] A. Kaltenbrunner, V. Gómez, and V. López. Description and prediction of slashdot activity. In *Proceedings of LA-WEB 2007*, 2007.
- [6] A. Kaltenbrunner and D. Laniado. There is no deadline - time evolution of Wikipedia discussions. In *Proceedings of WikiSym’12*. ACM, 2012.
- [7] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.
- [8] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *ICWSM-11 - 5th International AAAI Conference on Weblogs and Social Media*, 2011.
- [9] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of WWW2012*, pages 251–260. ACM, 2012.
- [10] N. Litvak and R. van der Hofstad. Scale-free graph sequences are not disassortative. *arXiv preprint arXiv:1202.3071*, 2012.
- [11] M. Mestyán, T. Yasseri, and J. Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *arXiv preprint arXiv:1211.0970*, 2012.
- [12] C. Okoli, M. Mehdi, M. Mesgari, F. Å. Nielsen, and A. Lanamäki. The people’s encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. Available at SSRN, 2012.
- [13] F. Ortega. *Wikipedia: A Quantitative Analysis*. PhD thesis, Universidad Rey Juan Carlos, Madrid, Spain, 2009. <http://libresoft.es/Members/jfelipe/phd-thesis>.
- [14] M. Osborne, S. Petrovic, R. McCreddie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using Twitter and Wikipedia. In *Proceedings of TAIA’12*, 2012.
- [15] J. Ratkiewicz, A. Flammini, and F. Menczer. Traffic in social media i: paths through information networks. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 452–458, 2010.
- [16] A. J. Reinoso, R. Muñoz-Mansilla, I. Herraiz, and F. Ortega. Characterization of the Wikipedia traffic. In *Proceedings of ICIW 2012*, pages 156–162, 2012.
- [17] B. Suh, G. Convertino, E. Chi, and P. Piroli. The singularity is not near: slowing growth of wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, page 8. ACM, 2009.
- [18] G. Szabo and B. Huberman. Predicting the popularity of online content. *arXiv preprint arXiv:0811.0405*, 2008.
- [19] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.
- [20] C. Wang, M. Ye, and B. A. Huberman. From user comments to on-line conversations. In *Proceedings ACM SIGKDD 2012*, pages 244–252, New York, NY, USA, 2012. ACM.
- [21] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.
- [22] T. Yasseri, R. Sumi, and J. Kertész. Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS ONE*, 7(1):e30091, 01 2012.

APPENDIX

Piecewise-linear approximation

Using the definition of the stages in the page view dynamics of an average promoted article in English Wikipedia we define a piecewise-linear approximation of the logarithm of the average number of views a promoted article attracts during the t -th hour as $g(t)$:

$$g(t) = a_t + b_t t,$$

where $a_t = 0$, $b_t = 7.6749$ for $t = 1$; $a_t = 7.6851$, $b_t = -0.0102$ for $t = 2, \dots, 24$; $a_t = 45.1967$, $b_t = -1.5103$ for $t = 25$; and $a_t = 6.2234$, $b_t = -0.0113$ for $t = 26, \dots, 95$.

Circadian patterns correction

Since the total page-view activity in Wikipedia varies during the day, these circadian cycles influence the promoted page popularity and caused the differences between the observed values and the approximation $\exp(g(t))$. Following [19] we first use a redistributed time scale where one hour equals $1/24$ of the daily traffic of the *Main Page* T . We note that this value could be a constant ($T_{main} = 6\ 050$ for our dataset) or change every day, month, or any time period. A new hour t' is therefore is the time interval which takes the *Main page* to accumulate from $\frac{t'-1}{24}T$ to $\frac{t'}{24}T$ views. This “decycling” allows us to ignore the dependency of the promoted article’s popularity on the total website traffic. In Figure 15 we plot the average number of views a promoted article attracts measured in the new time scale (Redistributed, $c=0$).

After applying the new time scale we mitigate the dependence of the promoted page views on the time of the day, however we did not remove it completely. A possible explanation would be that the normalization by T is not sensitive enough to the hourly changes in the traffic of the *Main page*. We propose to remove some constant fraction of the *Main page* view data before performing the redistribution in order to give more value to these traffic changes. Thus, we denote as $m(t)$ the average number of *Main page* views for given hour $t = 1, \dots, 24$. Recall that $T = \sum_{t=1}^{24} m(t)$ and define

new redistribution parameter T^* as follows:

$$T^* = \sum_{t=1}^{24} m^*(t) = \sum_{t=1}^{24} [m(t) - c \min_t m(t)],$$

where $c = \arg \min \left[\sum_{t^*=0}^{95} (\log(v_{t^*(c)}) - g(t^*))^2 \right]$ and $v_{t^*(c)}$ is the number of views of an average promoted page at time t^* in new time scale defined by T^* . In Figure 15 we plot the examples of the redistributions for different values of c . We find that $c = 0.406$ is the optimal value for “decycling” based on the *Main page* views. In other words one needs to remove 40% of minimum of the hourly traffic of the Wikipedia *Main page* to make an optimal correction of the circadian patterns.

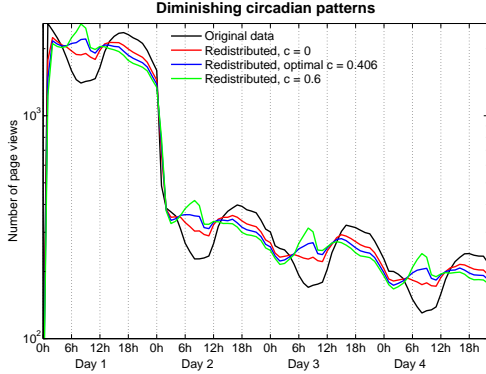


Figure 15: Comparing different parameter values for improving the estimate. (And the function $g(t)$ in the subplot)